Predicting Student Success based on their Stress with Deep Neural Neworks and Gradient Boosting

Mousa Aricat  
Department of Information and Tehnology  
Al-Hussein Technical UniversityAmman, Jordan  
Dr Raneem Qaddoura

# **Abstract**

This research investigates the impact of stress levels on academic performance among university students through using predictive models trained on a dataset. By discovering insights from experts and analyzing a Kaggle dataset, significant correlations emerge, highlighting stress as a key determinant. Preprocessing techniques prepare the data for predictive modeling, with gradient boosting showing robust performance, achieving an accuracy of 93.18% and similar metrics. The findings underscore the importance of addressing student stress and offer actionable recommendations for stakeholders. Future research will focus on refining predictive models and using innovative technologies to support student well-being and academic success.

# **Introduction**

In today’s data driven world, we are living in the era were enormous amounts of data are generated every minute, as there are now many different sources and technologies, such as social media applications, or sensors technology, that each generate those big amounts of data, leading us to enter the “Age of Data” (Nti, Quarcoo, Aning, & Fosu, 2022). The definition of data is that it is the collection of events and observations, leading that it is continuously increasing and getting more and more complex (Saeed & Husamaldin, 2021). This explosion of the digital information has given rise to the definition of Big Data, and the start of its era.

The term “Big Data” has spread into our lives that hearing it is not considered something new, whether we hear it in places like e-commerce, social media, or even research papers (Favaretto, De Clercq, Schneble, & Elger, 2019). It can have multiple definitions, as some define big data as the huge, enormous amounts of data, or the data that cannot be dealt with using normal methods, or maybe they define it based on the collection methods that have been used. Big Data has many different characteristics that defines it, these can be defined into 3, 5, 7, or even up to 10 different characteristics, as different research papers done on big data define them differently, these are called the Vs of big data (Saeed & Husamaldin, 2021). It is widely know that there are 5 main characteristics for the big data, which are: Volume (size of data), Velocity (the speed of data to be analyzed), Variety (the various data sources), Veracity (different values), and Value (the value the data gives). (Saeed & Husamaldin, 2021)

These days students are going through multiple factors that affect their level of performance, whether it was the ever-growing complexity of the educational ecosystem, or the mental state of the students. This has led to a challenge when it comes to identifying the key determinants to the students’ success. This research will apply Deep Neural Networks and Gradient Boosting models with the data that was collected on students, in order to find the things that make the student more successful.

This research will contribute in the following:

* Do analysis of interviews made with a psychologist and an instructor in the field of Artificial Intelligence.
* Develop a comprehensive analysis of a dataset that will have data for the students based on their study habits, stress levels, and other factors that are believed to influence the students’ success.
* Design and build, then train a Deep Neural Network and a Gradient Boosting model to predict the academic outcomes and performance based on the responses and observations from the survey.
* Provide actionable insights for the educations, administrators and institution based on the finding of the DNN and Gradient Descent models.

**Structure of the paper**  
The organization of this paper is as follows. In Section 2, the review of artificial neural networks (ANN) and Gradient Boosting is provided, along with a summary of its key aspects. Section 3 details the methodology employed for statistical testing, ANN and Gradient boosting models. Following this, Section 4 presents the results of statistical evaluations and performance metrics. Finally, Section 5 concludes the paper by summarizing the findings and their implications.

# **Literature review**

Within the range of investigating the intersection between the students’ stress factors and their success using big data and prediction models. This section provides a concise review of related studies and research papers to it. It also shows how predictive models are used in these fields and what are the different approaches to it.

Using machine learning models to predict the outcomes of the students’ performance was used amongst different research papers. (Lau, Sun, & Yang, 2019) used one ANN model with two hidden layers, one input and one output layers. (Ahujaa & Banga, 2019) focused on classification models and used four of them which were Logistic Regression (LR), Nayve Bayes (NB), Random Forest (RF) and SVM. (Hussain, Zhu, Zhang, Abidi, & Ali, 2018) and (Waheed, Hassan, Aljohani, Hardman, & Nawaz, 2019) combined the use of both machine learning and deep learning models and used ANNs, SVM, LR, and DT. As for (Imran, Latif, Mehmood, & Shah, 2019), they used ensemble method by using different models like the J48 classifier, then a Meta classifier to enhance the performance of J48. (Hooshyar, Pedaste, & Yang, 2019) used different models for both classification and regression features. (Ameen, Alarape, & Adewole, 2019) used many classification models including DT, LR, RF and more.

Data collecting and preprocessing is an essential step when doing machine learning models. (Lau, Sun, & Yang, 2019) collected data from 1000 students, but most of the dataset’s instances were females and the imbalance of the data was not handled. And for (Imran, Latif, Mehmood, & Shah, 2019), they have a dataset of 1044 and handled the imbalance of the dataset pretty well. (Ahujaa & Banga, 2019) had a smaller dataset of only 206 students, and they preprocessed the data by categorizing the labels. (Waheed, Hassan, Aljohani, Hardman, & Nawaz, 2019) had the biggest dataset from those papers with 32,593 students, they also did work on categorizing the labels based on different aspects.

When it comes to predicting the students’ success using those models, the results were based on different factors. For example, (Lau, Sun, & Yang, 2019) focused on the students’ performance on entrance exams and found the co relations and how would they affect the GCPA. (Imran, Latif, Mehmood, & Shah, 2019) had almost the same approach, they ranked the features that are most important like the absences rate, if the student failed before, what are their father qualification, etc. and added the features that ranked high to the models, then predicted the success of the students based on it. On the other hand, (Ahujaa & Banga, 2019) predicted only the stress levels of the students.

Model evaluation is an essential part to make sure that the models performed well, these models can be measures using different metrics such as accuracy for classification problems and R2 score or MSE for regression problems. (Lau, Sun, & Yang, 2019) achieved an accuracy of 84.8% using ANNs, which is not considered the best accuracy as the model had some limitations. (Ahujaa & Banga, 2019) achieved almost a similar accuracy with 85.71 using SVM. (Imran, Latif, Mehmood, & Shah, 2019) scored a considerably high accuracy using the ensemble methods and achieved 95.78%. (Hooshyar, Pedaste, & Yang, 2019) achieved the highest score amongst those paper with an accuracy of 96%, but the dataset was small. On the other hand, (Hussain, Zhu, Zhang, Abidi, & Ali, 2018) scored the lowest score with an accuracy of 80% using SVM.

When also speaking with the evaluation of the results, some of the papers only focused on the results of the model where others found some different relationships and measures that could help in the selected topic. Although (Lau, Sun, & Yang, 2019) did not achieve the highest accuracy, but co relations and statistical findings between different features with the label were given. On the other hand we have (Ahujaa & Banga, 2019) where they did not mention the important features mathematically that could affect the students’ performance. (Imran, Latif, Mehmood, & Shah, 2019) applied many measures that gave us information, but they were only related to the performance of the models.

Although these papers did provide good insights and work on predicting the students’ performance, not all of them took stress into consideration as some just took the information of past exams, redoing of lectures, or even their parents status. However, on this paper, stress is used as a main factor to see how it affects the performance of students, alongside other features and factors that will be collected from university students, to really check and see the important factors that affect their performance.

# **Research questions, hypotheses and objective.**

* What are the key determinants of the academic performance, identified through the DNN and the Gradient Boosting models that significantly contribute to students’ academic performance, and how do these determinants interact with one another?
* How effectively can Deep Neural Network (DNN) and a Gradient Boosting Models predict students’ academic success based on a comprehensive survey encompassing study habits, stress levels, and other relevant factors?

Hypotheses:

1. The predictive ability of the Deep Neural Networks and the Gradient Boosting model in determining students’ performance and academic success based on their study habits, stress levels, and other relevant factors, is statistically significant, indicating that stress levels play a significant role in predicting the student’s academic success.
2. The impact of stress levels on students’ academic performance is statistically significant, indicating a strong and measurable relationship between stress levels and academic success.

Objectives:

1. Evaluate the capability of Deep Neural Networks (DNN) and the Gradient Boosting model in predicting students’ academic success based on factors such as stress levels.
2. Quantify the impact of stress levels on student’s academic performance and determine the strength of the relationship between stress levels and academic success using statistical measures.

# **Data Description**

**4.1 Primary and Secondary Data.**

Primary Data is considered as the original data that is collected for the first time, it is usually done through personal experiences, or evidence, and is usually used for research purposes (IWH, 2015). Primary data can be collected using different resources, as usually researches collect the data themselves, they could use surveys, interviews, or even direct observations (IWH, 2015).

For the primary data in this research, I conducted two interviews, one with a psychologist (Dr Waleed Sarhan), and the other one with an AI instructor at Al-Hussein Technical University (Dr Rami Al Ouran). These two interviews covered different questions that helped in knowing the direction this research will be headed to.

Secondary Data can be defined as the data that has been already collected from other researches, for their purpose, and not collected for the purpose of this research (IWH, 2015). They can be accessed through different sources like the internet, or going to different organizations to get the data from them, or even ask other researches for their data (IWH, 2015). Secondary data is used to back up the research or even provide backup for the primary data (IWH, 2015).

For the secondary data in this research, I have found a dataset from Kaggle, the data set had some important information that will help this research, it was about the student stress factors as it made a comprehensive analysis about it.

**4.2 Primary Data Analysis:**

There were two interviews made, one with a famous psychologist in Jordan (Dr Waleed Sarhan) and with an instructor in the field of AI (Dr Rami Al-Ouran). The goal of these interviews was to ask questions that provide meaningful data to continue with this research and see how much the research questions are attainable. Another thing is that these interviews really helped in Also, each one of them was asked questions related to their field, and gave the wanted information regarding this research.

The first interview with Dr Waleed Sarhan had some main themes each with different codes that focuses on a specific topics or questions in the research, the reason these themes were used was to narrow the insights from the interview and get the data we want out of them. The themes and their codes were:

* **Theme 1:** Common Stressors among students:
  + **Codes**: Studying, University, Classes, Relationships.
* **Theme 2:** The impact of stress on their academicperformance.
  + **Codes**: Sleep, Appetite, Mood, following classes.
* **Theme 3: Factors contributing to their stress:**
  + **Codes**: Personal life, family, expectations, abilities
* **Theme 4: Variations in Stress Levels:**
  + **Codes**: Expectations, Class load, stress, family.
* **Theme 5: How common is Stress among University Students:**
  + **Codes**: common, stressors, university students.
* **Theme 6:** Coping Mechanisms for University Students with Stress.
  + **Codes**: Coffee, Tea, Smoking, Power Drinks, Addictive substances, Friends.

In summary, our thematic analysis revealed several key insights regarding the factors influencing student stress and coping mechanisms. Firstly, for Theme 1, students commonly encounter stressors related to academic demands, including studying, university classes, and interpersonal relationships. Secondly, as highlighted in Theme 2, stress has a significant impact on students' academic performance, affecting their engagement in classes, completion of assignments, and overall well-being. Thirdly, as indicated in Theme 3, various factors contribute to students' vulnerability to stress, encompassing personal experiences, family dynamics, academic expectations, and time management skills. Moreover, for Theme 4, the nature and intensity of stress vary among students, influenced by factors such as parental expectations and academic rigor. Additionally, as observed in Theme 5, stress prevalence is pronounced among students facing academic challenges or pursuing undesirable majors. Lastly, for Theme 6, students employ diverse coping strategies, including unhealthy habits like smoking and substance abuse, to alleviate stress. These findings underscore the multifaceted nature of student stress and highlight the importance of targeted interventions to support their well-being and academic success.

As for the second interview which happened to be with Dr Rami Al Ouran, the answers were divided regarding several themes, and based on those themes and codes, the needed insights were captured and used in the research to go on with the work. The themes and codes are the following:

* **Theme 1:** Deep Neural Networks in Predicting using Big Data. **Codes**: Predicting, Complex data, features, results.
* **Theme 2:** Deep Neural Networks in Predicting the Students’ academic success based on Stress. **Codes**: relations, students, features, performance, accurate results, data.
* **Theme 3:** Mathematical Evidence on Stress and Academic Performance. **Codes**: Mathematical, evidence, Coefficients, relationships.
* **Theme 4:** Recommendations for other Methods and Models than Deep Neural Networks. **Codes**: Small Data, Overkill, Regression problem, Linear, models.
* **Theme 5:** Importance of Explainable AI with Deep Neural Networks. **Codes**: Explainable AI, complex models, results, explain.

In examining the role of Deep Neural Networks (DNNs) in predicting academic success and understanding student stress, several key insights emerge. Firstly, as highlighted in Theme 1, DNNs demonstrate effectiveness in handling complex datasets with numerous features, making them well-suited for predicting outcomes using big data. Their ability to navigate intricate data structures and extract meaningful patterns contributes to the generation of accurate results. Secondly, for Theme 2, DNNs emerge as a valuable tool for forecasting students' academic achievements based on factors such as stress levels. By uncovering intricate relationships between stress and academic performance, DNNs offer reliable insights into student success, leveraging their robust data processing capabilities. Thirdly, in Theme 3, DNNs exhibit the capacity to provide transparent mathematical evidence elucidating the direct impact of stress on students' performance and GPA. Through meticulous data training, DNNs yield coefficients that delineate the influence of each feature on academic outcomes, shedding light on their interplay. Additionally, as discussed in Theme 4, the suitability of DNNs is contingent upon dataset size and complexity. While DNNs excel in handling large datasets, their utility may diminish when confronted with smaller, less complex datasets, warranting consideration of alternative models like linear regression for optimal performance. Lastly, Theme 5 underscores the importance of Explainable AI (XAI) in elucidating the workings of complex models like DNNs. Given the inherent complexity of DNN architectures, XAI plays a crucial role in demystifying model outputs and enhancing interpretability, enabling stakeholders to grasp the underlying mechanisms driving predictions. These insights underscore the versatility and significance of DNNs in predictive analytics while emphasizing the need for interpretability and model transparency afforded by XAI techniques.

**4.3 Secondary Data Analysis**

The analysis of the secondary data from the Kaggle dataset provided valuable insights into the relationship between stress levels and academic performance among university students. By exploring the dataset, we uncovered patterns and correlations that shed light on how different factors, such as stress, sleep quality, and anxiety levels, impact students' success. Through visualizations and correlation analysis, we identified stress levels as a significant predictor of academic performance, with higher stress levels correlating with lower grades. Additionally, the dataset revealed other important variables, such as bullying, teacher-student relationships, and sleep quality, which showed strong correlations with academic outcomes. This analysis helped us understand the complex interplay of various factors influencing student success and informed the development of our predictive models.

**4.4 Ethical Issues.**

During the data collection process, several access and ethical issues were encountered. Firstly, getting the permission from the participants, which involved navigating and seeing if there should be any regulations we should stick to. Secondly, ensuring participant confidentiality and privacy was protected, particularly when collecting sensitive information such as stress levels and academic performance.

Another point is using a dataset from an onlince source required making sure to use it in an appropriate manner and for research purposes.

To address these issues, some steps were followed, like getting the permissions from the people that the interview will be done with, and making sure that they are okay in using their information in this research. As for the secondary data which is the dataset from Kaggle, we made sure that it is okay to use it and can be used for research purposes.

**4.5 Merits and Limitations.**

The data collection process covered different aspects. By doing two interviews, we had a bigger understanding of the students’ stress factors and how it impacts their success, and how can we utilize AI with it. By directly engaging with experts in psychology and academia, detailed information were gained on the relationship between stress levels and academic performance of the students, as well as the various coping mechanisms adopted by students. The insights gained from the interviews guided the research direction and helped refine the research questions to focus on the most important issues. For instance, understanding the common stressors among students and the allowed us to tailor our investigation to explore these factors in depth. As for the secondary data, the dataset that was taken from Kaggle had valuable information that gave us what we need in terms of insight and valuable patterns. The secondary dataset facilitated comparative analysis across different groups of students based on various demographic and academic characteristics. By comparing academic performance and stress levels among different students, we gained insights into differences and inequalities in educational outcomes, informing targeted interventions and policy recommendations.

For the limitations of the primary data, some of the answers given by the interviewees were only briefly and they could not provide any more details. Also, some questions came up after the interviews were done, this meant that the information we could have gathered from this question was gone as we could not interview them again due to their busy schedules. Also, the sample size of the interviews was relatively small, involving only two experts in psychology and AI instruction. This limited the diversity of perspectives and may not have captured the full range of experiences and insights related to student stress. Finally, the interviews may not have covered all relevant aspects of stress and academic performance, potentially overlooking important factors that could influence the findings

For the secondary data, one limitation was the potential for response bias. Another limitation was that the dataset size was small and had about 1000 observations, which may not be enough to find the patterns we need and train the models on. Also, there were not any research papers on this dataset which made us do all the work of analyzing it. And finally, only two features related to the stress factors were mentioned, and the other features were related to different aspects of the students’ lives.

**4.6 Tools for the data collection process.**

There were different tools that we used, as for us to collect the qualitative primary data we used interviews with two experts in the field of this research and which they provided the valuable information we could use in this research to get the best outcome. And as for the dataset, it was collected using surveys, but since it is a secondary data source, we only took it from Kaggle with permission of course.

And in the terms of the data analysis, some tools were used like the pandas library in python, also the seaborn and matplotlib as these three libraries helped in analyzing the data, finding the patterns that provided the information we need for this research, and seeing the insight we need to make sure that the dataset is a good match and can be used for this research to then carry on and apply the models, as every pattern and insight in the data can indicate a different meaning.

# **Methodology**

The methodological approach of this research follows Saunders' research onion, guiding the philosophical, theoretical, and methodological aspects. At the core, an interpretivist stance is taken to explore the subjective nature of predicting academic success from stress levels. Inductive reasoning is employed to generate insights from qualitative data gathered through interviews. Qualitative methods are chosen to delve deeply into students' experiences. A cross-sectional design is adopted to capture a snapshot of students' experiences. These methodological choices form a systematic approach, aiding in achieving research objectives and gaining valuable insights into stress and academic success.

**Research Philosophy (Inner Layer)**

The first layer of the research onion involves selecting a research philosophy, which encompasses positivism, interpretivism, and realism. For this study on predicting academic success based on stress levels, an interpretivist approach was chosen. This is because we aim to understand the subjective experiences and interpretations of students regarding stress and its impact on their academic performance. Interpretivism allows for exploring the complexities of human behavior and the context in which it occurs, which is essential for understanding the nuanced relationship between stress and academic success.

**Research Approach**

The next layer involves selecting a research approach, which includes deductive and inductive approaches. Given the exploratory nature of our study and the need to generate new insights, an inductive approach was considered appropriate. By collecting qualitative data through interviews and then deriving patterns and themes from the data, we can develop hypotheses or theories about the relationship between stress and academic performance. This aligns with an inductive approach, which allows for theory-building based on empirical observations.

**Research Strategy (Outer Layer)**

At this layer, the focus is on selecting specific research strategies such as experiments, surveys, or case studies. Considering our research objectives and the need to gain insights from real-life experiences, a qualitative research strategy was chosen. Qualitative methods like interviews enable us to delve deeply into students' perspectives, experiences, and information regarding stress and academic performance. Through open-ended questions, we can uncover rich data that provides a better understanding of the study.

**Choices of Methods**

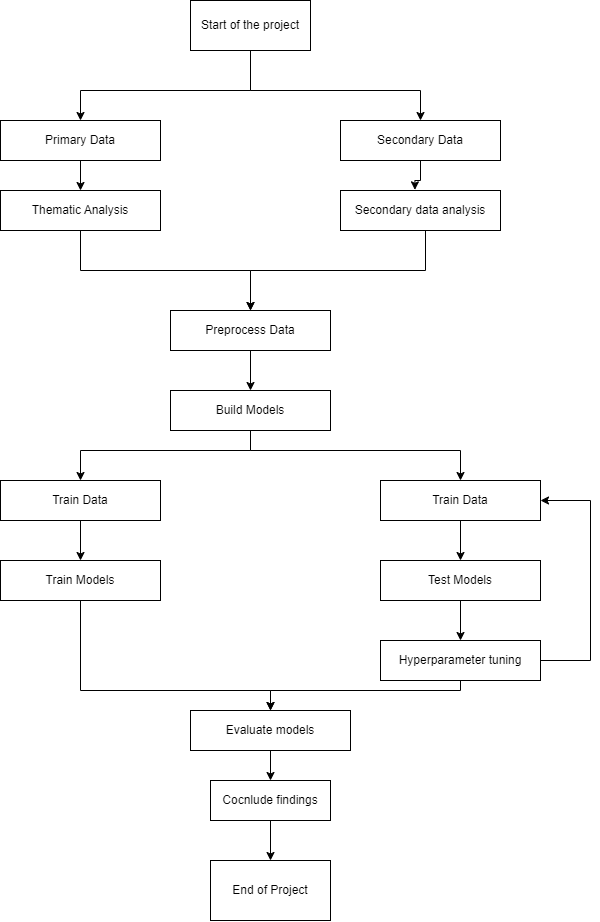
Careful decisions were made regarding data collection techniques, sampling methods, and data analysis approaches to ensure the robustness and reliability of the findings. This corresponds to the choices of methods layer in Saunders' research onion, wherein researchers select specific methods to address their research objectives. By sticking to a framework, the research aims to provide a comprehensive understanding of the relationship between stress and academic success among university students.

**Time Horizon**

Given the relatively short duration of our study and the focus on understanding current experiences and perceptions of students, a cross-sectional design was chosen. This allows us to capture a snapshot of students' stress levels and academic performance at a specific point in time, without the need for longitudinal data collection.

**Techniques of Data Collection (Outer Layer)**

In this layer, we determine the specific methods for collecting data that align with our research strategy and approach. Given our qualitative research strategy, we opt for techniques such as semi-structured interviews to gather rich, detailed insights from university students regarding stress and academic performance. Semi-structured interviews offer flexibility in probing and exploring participants' responses, allowing us to delve deeply into their experiences and perspectives.

**Flow Chart**

**Start of the project**

**Primary Data:** Conducting interviews with relevant parts, such as educators, or experts, provides valuable firsthand insights into the research topic. These interviews help in understanding perspectives, identifying key themes, and validating findings from other data sources. They give us the information that we may be unsure about and any important data that will give us relevant and important insights to continue with the research.

**Thematic Analysis:** Thematic analysis of interview data involves identifying patterns, themes, and recurring topics within the responses. This process helps in uncovering insights, understanding participant perspectives, and generating qualitative data that will help in the research, the way it is done by doing themes and for each theme there will be codes generated, and each theme will cover a specific topic that will help in getting the wanted insight or pattern from the interview. This is considered one of the essential steps in the data collection and processing process as this will help in getting patterns and insights, and without it the work with the interviews will be considered so much harder, and it will take so much work to get any insights from interviews.

**Secondary Data:** Identifying and acquiring secondary datasets that align with the research objectives is essential for augmenting primary data and providing additional context or validation. For example, in our study, we might seek datasets containing information on stress levels, academic performance, and related factors among students. Having secondary data is essential, and the reason for this is that the primary data is not always enough, and sometimes it needs to be backed up by secondary data. Also, it saves time to get secondary data that is already there, especially if it was collected for similar reasons.

**Secondary Dataset Analysis:** Analyzing secondary datasets involves exploring the data to identify trends, correlations, and patterns relevant to the research objectives. This step provides additional insights into the phenomenon under study and enriches the overall analysis. Analyzing this data can be using different tools and methods, like some famous libraries in python such as matplotlib and seaborn. This step is important in the research because with it we will be able to identify and analyze the features, know more about the dataset, and what are the important things that should be taken into consideration. Also, analyzing it will help in the preprocessing step ahead, as we will have an idea of the statistical measures, the null counts, duplicates, and many other things that will bring value to the table.

**Preprocess Data:** Data preprocessing involves cleaning and transforming raw data to ensure its quality and suitability for analysis. This step includes handling missing values (nulls), removing duplicates, scaling features to a consistent range, and addressing class imbalance through techniques like SMOTE oversampling.This step is also one of the most important steps, especially in this research, as datasets are usually not clean and need to be handled in different ways, which takes work and time. This is where preprocessing comes in to help with making the data ready for the next steps, it involves different steps that will make the data the most suitable for the training and testing, thus making this step as one of the most important.

**Build Models:** Constructing predictive models involves applying machine learning algorithms to the preprocessed data to forecast outcomes or identify patterns. In our study, we utilize advanced techniques such as Deep Neural Networks and Gradient Boosting to develop models capable of predicting students' academic success based on stress levels and other factors. They will be applied using the scikit learn library which applies these models at ease without much work, and we can hyper tune them easily. This is important as the research is based on making predictive models predict the outcome of the students’ performance based on their stress, so they are like the main block of this research which makes this step essential.

**Train Data:** this is where we get the data that will be used for the train ready. And for this case since there is one data set, the training data will be part of the original data and not a separate dataset. The reason for the division is that the model will not see all the data and overfit, and then we can test it to see how good it did. The selection of appropriate train data is crucial for ensuring the model's ability to generalize well to unseen data. Researchers may employ techniques such as random sampling or cross-validation to split the dataset into train and test subsets, which balances the distribution of classes and makes sure they have data to represent samples for training.

**Train Models:** Training the predictive models involves feeding them with labeled data and iteratively adjusting their parameters to minimize prediction errors. This process enables the models to learn from the data and improve their predictive accuracy over time. Training models is essential as it allows us to leverage machine learning algorithms to analyze the train data and uncover meaningful insights. This step will get the models to be ready for testing and then deployment.

**Test Data:** test data is the other part of the data left from the division for the training, it is usually much smaller than the training data, and it is used to see how good the model did really. It is the data that the model has not seen at all, which makes it a challenge for the models to predict this data as they have not trained on it. The test data step also plays a crucial role in validating the performance of the trained models and assessing their ability to generalize to new data.

**Test Models:** Testing models involves applying trained machine learning models to the test data to generate predictions and evaluate their performance. This process entails feeding the test data into the trained models and comparing the predicted outcomes against the actual labels to measure predictive accuracy. We can assess the models' performance using various evaluation metrics and visualization techniques to gain insights into their strengths and weaknesses. Testing models provides valuable feedback on the efficacy of the algorithms and informs decisions regarding model selection, parameter tuning, and refinement. Additionally, researchers may conduct comparative analyses between different models to identify the best-performing approach for the given task.

**Evaluate Models:** Evaluating models was a critical step in assessing their performance and determining their effectiveness in addressing the research objectives. By employing various evaluation metrics such as accuracy, precision, recall, and F1 score, we gained insights into how well the trained models performed in predicting academic success based on stress levels and other relevant factors. Additionally, techniques like cross-validation helped to validate the robustness of the models and ensure their reliability across different datasets. This phase allowed me to compare different models, identify strengths and weaknesses, and make informed decisions regarding model selection and refinement. Ultimately, the evaluation of models provided valuable insights into their predictive capabilities and their potential to contribute to the advancement of research in the field of student well-being and academic success.

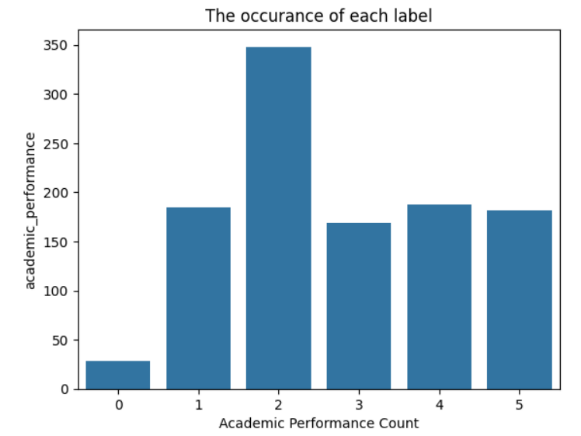
**Conclude Findings:** Drawing conclusions based on the research findings involves summarizing key insights, addressing research questions, and discussing implications for theory, practice, and future research. The conclusion of the study provides closure to the research process and highlights its contributions to the field.

**End of Project**

# **Analysis and Results**

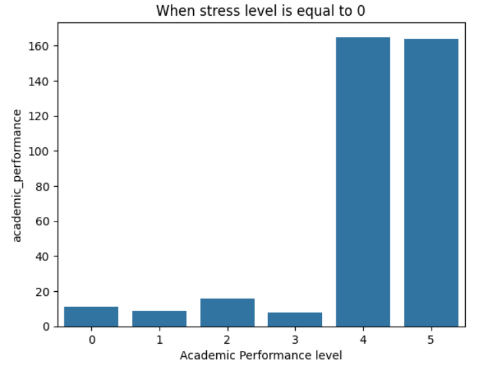
**6.1 Analysis of the Dataset.**

The dataset that we used had a total of 20 features and 1 label (the academic performance), the academic performance had 6 values in it which ranged from 0 to 5 on how well the student is doing academically. The figure below shows how many each of these values occurred.



It is clearly shown that the dataset is imbalanced on some labels like the 2 and the 0, so this is a thing that should be taken into consideration before training the model on this data.

The main purpose of this research was to see how much the stress factor specifically affects the student’s performance, so in this dataset, there were three levels for the stress factor, 0 being the lowest with no stress, 1 being the middle with somewhat stress, and 2 being the highest stress level. So to find out the relationship between them and how the academic performance is affected, we used a bar plot to plot the observations of the label (academic performance) with each stress level.



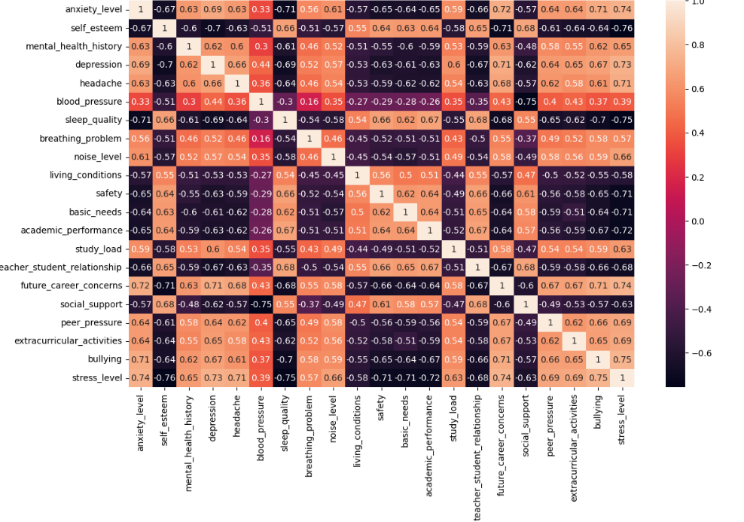
A graph of stress level

Description automatically generated

**A graph of stress level

Description automatically generated**

From those figures, we can get the insights that the stress factor is an essential factor that affects the academic performance in a huge way. Looking at the first figure when the stress level was at 0 (the lowest) we can see that most of the students with it got the highest performance (4,5). Moving on to when the stress was at the neutral level we got the insight that the students performed moderately getting most of the scores at (2,3). And finally, when the stress was at the highest level, we can see that the students got the lowest scale in terms of academic performance (1,2). These insights give us a big sign of how important the stress factor is related to students. And to get a whole view on what are the features that affect the students academic performance, we decided to plot the co relation table.



From it we also got the data we needed that the stress factor is indeed the highest co related feature with the label, with a negative (if stress increases the performance decreases) co relation as -0.72.

Looking at the co relation table we even found more features that have strong co relations, but we will stick to only write the co relation and not visualize them as the main purpose of this research is to see how the academic performance is affected by the stress specifically.

Some of these features are the following: **Bullying** with a negative co-relation of -0.67 making it rank 2nd in terms of the highest co-relations alongside the **Teacher Student Relationship** with a positive co-relation of 0.67, which means that the more students are being bullied, the lower their academic performance will be, and the more they are in good terms with their teacher, the higher their scores will be. Another important feature that also ranked second was the sleep quality, as it got a co-relation of 0.67 which makes sense that the more sleep a student gets the higher likelihood of them succeeding in their performance. And ranking third we have the **anxiety level** with a negative co-relation of -0.65, which means that if a student has high anxiety levels, it will affect their marks negatively.

And many other features showed high co-relations, which was one of the properties about this dataset, that each feature can affect the academic performance and is considered important.

**6.2 Preprocessing the data.**

After collecting the data, the next step was to prepare it for analysis. This involved preprocessing the data to ensure it was suitable for training the model (gradient descent). There were no nulls in the dataset and no duplicates which saved us some time in preprocessing it. The preprocessing steps included using techniques like standard scaling and SMOTE oversampling. Standard scaling helped in standardizing the features of the dataset by subtracting the mean and dividing by the standard deviation. This ensures that all features have a similar scale, which can improve the performance of the model. SMOTE oversampling is a technique used to address class imbalance in the dataset. In our case, it helps to balance the distribution of classes related to stress levels, ensuring that the model is trained on a more representative sample of the data. By applying these preprocessing techniques, we were able to enhance the quality of the dataset and prepare it for training the predictive model. This step is crucial in ensuring that the model can effectively learn from the data and make accurate predictions about student success based on their stress levels and other relevant factors.

**6.3 The results of the model.**

We initially tested a Deep Neural Network (DNN) model to assess its predictive capabilities on our dataset. The model had one input layer with 64 neurons, one hidden layer with 32 neurons, and an output layer. Our initial experimentation with the DNN yielded a modest accuracy of only 40%. This outcome underscored the need for further refinement and optimization to enhance the model's predictive power. It also highlighted the iterative nature of our approach, wherein we iteratively tested different methodologies to find the most effective modeling technique for our specific research objectives and dataset characteristics. This initial experimentation with the DNN showed us that they might not be the best choice as they are usually so complicated and could be shown as an overkill since the dataset is so small and does not have that many observations for the neural network to learn from. And even before getting the 40% accuracy on the testing data, the model gave an accuracy of 85% in the training data, which means that the model was overfitting due to its complexity and the small size of the dataset.

So, we decided to continue with the gradient boosting, and after running grid search with gradient boosting to fine-tune our model and find the best hyperparameters, we found the best settings to be a total of 250 estimators (basically, the number of decision trees used in the model), with each tree having a maximum depth of 3 (meaning how many layers deep each decision tree can go), and a learning rate of 0.1 (which controls how much each tree contributes to the final prediction). With these settings, we tested our model on the dataset to see how well it performed.

The results were impressive especially having the factor that the dataset is so small. Our model showed an accuracy of about 93.18%, which means it correctly predicted the outcome for around 93.18% of the students in our dataset. The precision, which tells us how often the model is right when it predicts a student will succeed, came out to be around 93.21%. When it came to recall, indicating how many of the students who actually succeeded were correctly identified by our model, it was also high at 93.18%. And finally, the F1 score, which is like a blend of precision and recall, giving us a balanced view of the model's performance, was calculated to be approximately 93.17%.

So, what does all this mean? Well, it suggests that our model is doing a pretty good job at predicting whether a student will succeed based on their stress levels and other factors we considered. These results give us confidence that the model is effective and could be useful in helping educators identify students who might need extra support to thrive academically.

**6.4 Examinations of the results and the findings.**

When examining the results of our two models’ performance, several key insights emerge that we need to take a good look at and consider. For the DNN, we can clearly see that it was not the best choice, and there are many reasons for this, going from its complex nature as it is defined as Deep Learning, and works usually better with complex problems, not simple ones with a simple dataset that is highly co-related like this one, so using a DNN is an overkill, which made us move to the next model, the gradient boosting. Firstly, the high accuracy, precision, recall, and F1 score in the test data using gradient boosting indicate that our model is robust and effective in predicting student success based on stress levels and related factors. This suggests that the features we selected, including stress levels, study habits, and other relevant factors, indeed play significant roles in determining academic outcomes, and that what was shown in the high co-relations they had was actually real and that this data is highly co-related which resulted in high results while having a small dataset.

Additionally, the consistency across multiple evaluation metrics reinforces the reliability of our model. The fact that accuracy, precision, recall, and F1 score are all within close proximity suggests that our model is well-balanced and does not exhibit significant bias towards any particular metric.

Furthermore, the chosen hyperparameters (number of estimators, maximum depth, and learning rate) seem appropriate for our dataset, as they result in a model with high predictive performance. This indicates that the grid search process effectively identified optimal settings for our gradient boosting algorithm.

However, while these results are promising, it's important to acknowledge potential limitations. For instance, the dataset used for training and evaluation may not fully capture the complexity of real-world scenarios, and there may be other unmeasured variables that also influence student success. Additionally, the performance of the model may vary when applied to different populations or contexts.

Overall, the examination of the results highlights the potential of our model to provide valuable insights into student success and inform targeted interventions to support students in their academic journey. Further validation and refinement of the model through additional testing and real-world application will be essential to maximize its utility and impact.

# Conclusion and future work

In conclusion, this study show the influence of stress on academic performance among university students. Through analysis and methodology, compelling insights were uncovered into the interplay between psychological well-being and achievement. The results of our predictive models, with accuracy rates reaching 93.18%, precision of 93.21%, recall of 93.18%, and an F1 score of 93.17%, underscore the efficacy of our approach in discerning patterns and making informed predictions. These numerical outcomes not only validate the significance of stress as a pivotal factor but also emphasize the potential of machine learning techniques in deciphering complex educational phenomena. Moreover, our findings underscore the imperative for proactive interventions and comprehensive support mechanisms to address student stress and promote holistic well-being. As we chart the course for future research, incorporating longitudinal studies and expanding the scope of predictive models, we remain committed to advancing our understanding of student success and fostering environments conducive to academic excellence and flourishing.

To address our research questions, we embarked on a systematic approach. First of all, we compiled a comprehensive dataset which shows different features of student life, including study habits, stress levels, and academic achievements. Subsequently, we undertook data preprocessing steps to ensure the dataset's cleanliness and suitability for analysis. Leveraging sophisticated machine learning methodologies, specifically the Deep Neural Network (DNN) and Gradient Boosting models, we delved into the data to unearth significant insights.

In response to the first question which was : “What are the key determinants of the academic performance, identified through the DNN and the Gradient Boosting models that significantly contribute to students’ academic performance, and how do these determinants interact with one another?”, the dataset was subjected to thorough exploration and preprocessing to ensure its suitability for modeling. Employing advanced analytical techniques, we discovered patterns and correlations among different variables. Through iterative experimentation with the DNN and Gradient Boosting models, we identified key factors that wielded substantial influence on academic success. These models provided invaluable insights into the intricate interplay between various determinants and their impact on students' academic outcomes, which means that this questions was successfully answered.

For the second research question, “How effectively can Deep Neural Network (DNN) and a Gradient Boosting Models predict students’ academic success based on a comprehensive survey encompassing study habits, stress levels, and other relevant factors?”, the effectiveness of DNN and Gradient Boosting models in predicting academic success was evaluated based on their performance metrics. These metrics included accuracy, precision, recall, and F1 score, which provided insights into the models' predictive capabilities. Through rigorous testing and comparison, the predictive performance of both models was assessed, allowing for a comprehensive understanding of their effectiveness in predicting students' academic success based on the surveyed factors.

Overall, the attainment of the research questions was facilitated by a methodical approach that involved data preprocessing, model training, and evaluation of model performance. Through this process, key determinants of academic performance were identified, and the predictive capabilities of DNN and Gradient Boosting models were assessed, contributing to a deeper understanding of student success factors.

Based on the findings of this research, it is recommended to explore diverse methodologies to further refine predictive models and deepen our comprehension of student performance. Firstly, expanding the selection of machine learning algorithms beyond those in this research, such as considering ensemble techniques like Random Forests or other models like Support Vector Machines, could provide insights into alternative modeling strategies and their efficiency in forecasting academic results based on stress levels and related features. Also, increasing the dataset size by getting more representative data from a wider range and variety of universities and educational places would give a more comprehensive base for analysis and model training and testing. This expansion may involve having other features beyond stress levels, such as socioeconomic background, extracurricular participation, and familial support, to capture a bigger view of the factors impacting student success and performance. Another thing is, trying to construct a dataset make specifically to the research, this would allow for more specificity in data collection and facilitate the exploration of associations between variables. Enhancing and refining the predictive models could help in integrating advanced techniques such as feature engineering, dimensionality reduction, and hyperparameter tuning to optimize model performance and generalization capabilities. Furthermore, utilizing explainable artificial intelligence methods to augment model interpretability and transparency would enable stakeholders to comprehend how models formulate predictions and empower them to make well-informed decisions regarding student support strategies. By embracing these recommendations and continuing to innovate in research methodologies and modeling approaches, educational stakeholders can propel the field forward and pave the way for more effective interventions to foster student well-being and academic achievement.

Looking forward, our next steps involve improving and broadening our research efforts. One big part of this is gathering our own dataset. We want to do this ourselves so that the data we collect fits exactly what we need for our research We plan to talk directly with students by doing surveys or interviews, but for the dataset it will be mostly surveys, these surveys will be directed for university students across all Jordan, and we might also consider doing one for the school students as they also face stress. This way, we can learn more about different things that might affect how well students do in university or even school. We also want to try out different types of machine learning models, like try to with the deep neural networks and try to make them work since we will most likely have a bigger and more complex dataset.

Looking ahead, there are several aspects for future research that could build upon the recommendations provided earlier in this study. Firstly, expanding the dataset size by collecting data from a larger and more diverse population of university students across different demographics and academic aspects would enhance the generalizability and robustness of predictive models. Additionally, incorporating a wider range of features beyond stress levels, such as study habits, social support networks, and lifestyle factors, could provide a more comprehensive understanding of the multifaceted influences on student academic performance. Moreover, refining predictive models with advanced techniques such as ensemble learning, neural architecture search, and model interpretability methods could improve model accuracy and facilitate the identification of actionable insights for stakeholders. Furthermore, exploring the feasibility of longitudinal studies to track students' academic trajectories and well-being over time would offer valuable insights into the dynamic nature of student success factors. Additionally, efforts to develop explainable artificial intelligence (XAI) techniques could enhance transparency and interpretability in predictive models, thereby fostering trust and acceptance among educators and students. Finally, establishing collaborative partnerships between academia, industry, and mental health organizations could facilitate the co-design and implementation of evidence-based interventions tailored to address the diverse needs of students. By pursuing these future research directions, scholars can contribute to the advancement of predictive modeling in education and the development of effective strategies to support student well-being and academic success in university settings.

# References

Ahujaa, R., & Banga, A. (2019). Mental Stress Detection in University Students using Machine. *Procedia Computer Science*, 5.

Ali, S. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to. *elsevier*, 52.

Ameen, A. O., Alarape, M. A., & Adewole, K. S. (2019). STUDENTS’ ACADEMIC PERFORMANCE AND DROPOUT PREDICTIONS: A REVIEW. *Malaysian Journal of Computing*, 26.

Favaretto, M., De Clercq, E., Schneble, C. O., & Elger, B. S. (2019). What is your definition of Big Data? 20.

Hooshyar, D., Pedaste, M., & Yang, Y. (2019). Mining Educational Data to Predict Students’ Performance through Procrastination Behavior. *Entropy*, 24.

Hussain, M., Zhu, W., Zhang, W., Abidi, S. M., & Ali, S. (2018). Using machine learning to predict student difficulties. *Springer Nature 2018*, 27.

Imran, M., Latif, S., Mehmood, D., & Shah, M. S. (2019). Student Academic Performance Prediction using. 13.

IWH. (2015, November). *IWH*. Retrieved from IWH: https://www.iwh.on.ca/what-researchers-mean-by/primary-data-and-secondary-data

Langer, M. (2021). What do we want from Explainable Artificial Intelligence. *Science Direct*, 24.

Lau, E. T., Sun, L., & Yang, Q. (2019). Modelling, prediction and classifcation of student academic . 10.

Nielsen, M. (2017). *Neural Networks and Deep Learning.* Springer Link.

Nti, I. K., Quarcoo, J. A., Aning, J., & Fosu, G. K. (2022). A mini-review of machine learning in big data analytics: Applications, challenges, and prospects. *A mini-review of machine learning in big data analytics: Applications, challenges, and prospects*, 17.

Robert-McComb, J. J. (2001). *Eating Disorders in Women and Children.* London: CRC Press.

Saeed, N., & Husamaldin, L. (2021). Big Data Characteristics (V’s) in Industry. *Iraqi Journal of Industrial Research, Vol. 8, No. 1*, 9.

Waheed, H., Hassan, S.-U., Aljohani, N. R., Hardman, J., & Nawaz, R. (2019). Predicting Academic Performance of Students from VLE Big Data using Deep. 34.

York, T. T., Gibson , C., & Rankin , S. (2015). Defining and Measuring Academic Success. *Practical Assesment, research and evaluation*, 21.